Maximum a posteriori covariance estimation using a power inverse wishart prior

Nielsen, Søren Feodor; Sporring, Jon

Publication date:
2012

Document version
Early version, also known as pre-print

Maximum A Posteriori Covariance Estimation Using a Power Inverse Wishart Prior

Søren Feodor Nielsen[a,*], Jon Sporring[b]

[a] *Copenhagen Business School, Solbjerg Plads 3, DK-2000 Frederiksberg, Denmark*
[b] *eScience center, Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark*

Abstract

The estimation of the covariance matrix is an initial step in many multivariate statistical methods such as principal components analysis and factor analysis, but in many practical applications the dimensionality of the sample space is large compared to the number of samples, and the usual maximum likelihood estimate is poor. Typically, improvements are obtained by modelling or regularization. From a practical point of view, these methods are often computationally heavy and rely on approximations. As a fast substitute, we propose an easily calculable maximum a posteriori (MAP) estimator based on a new class of prior distributions generalizing the inverse Wishart prior, discuss its properties, and demonstrate the estimator on simulated and real data.

Keywords: Covariance estimation, Bayesian method, maximum a posteriori, inverse Wishart distribution, Tracy-Widom distribution

1. Introduction

The problem of estimating a large covariance matrix with limited amounts of data occurs in many different applications of statistics such as image analysis, functional data analysis, quantitative finance, analysis of microarray data etc. We became interested in this problem through the study of shape variations in medical applications, e.g. X-ray images of human vertebra [3].

*Corresponding author
Email address: sfn.mes@cbs.dk (Søren Feodor Nielsen)

To study the shape variation in such data, images are annotated by a medical expert, and in the case of the vertebra 50 anatomically meaningful points were set on each 2 dimensional X-ray image, such that each shape is represented by a 100 dimensional vector. For such a high-dimensional space, the standard ML covariance matrix estimate requires in the order of 1000 annotated images to be of reasonable accuracy. Unfortunately, this is rarely available, since the annotation task is laboursome and medical experts are a limiting resource. Therefore, we have been looking into improved estimates for small samples of high dimension.

In this paper we propose a maximum a posteriori (MAP) estimator for the unknown covariance matrix based on a new class of prior distributions, which we call the power inverse Wishart distributions. We introduce the distributions in section 2 and derive the MAP estimator in section 3. We compare its properties with those of the usual inverse Wishart MAP estimator in section 4, derive some asymptotic results in section 5, and demonstrate its applicability on simulated (section 6) as well as on real data (section 7).

2. The Power Inverse Wishart Distribution

We start by defining a class of distributions on the set of positive definite $p \times p$-matrices. This class generalizes the well-known inverse Wishart distribution and, as we will argue in the following section, leads to tractable MAP estimators of an unknown covariance matrix of a multivariate normal distribution.

Definition 1. *The power inverse Wishart distribution with parameters $(\boldsymbol{\Psi}, m, q)$, where $\boldsymbol{\Psi}$ is a positive definite $p \times p$-matrix, $m \geq p$, and $q \in \{1, 2, \ldots\}$, is the distribution on the set of positive definite $p \times p$-matrices with density given by*

$$\mathcal{W}^{-q}(\boldsymbol{B}|\boldsymbol{\Psi}, m) = \frac{1}{c_{m,q}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\left(\boldsymbol{\Psi}^{-1/2}\boldsymbol{B}\boldsymbol{\Psi}^{-1/2}\right)^{-q}\right)\right) \frac{|\boldsymbol{\Psi}|^{qm/2}}{|\boldsymbol{B}|^{qm/2+p/2+1/2}}, \tag{1}$$

where $c_{m,q}$ is a normalization constant given by

$$c_{m,q} = \int \exp\left(-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{B}^{-q}\right)\right) |\boldsymbol{B}|^{-qm/2-p/2-1/2}\, d\boldsymbol{B}, \tag{2}$$

where the integral is over the set of positive definite $p \times p$-matrices.

The distribution is well-defined, when the integral in (2) is finite; we show this in the following theorem. For $q = 1$, the power inverse Wishart distribution is the well-known inverse Wishart distribution with density

$$\mathcal{W}^{-1}(\boldsymbol{B}|\boldsymbol{\Psi}, m) = \frac{|\boldsymbol{\Psi}|^{m/2} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}\boldsymbol{B}^{-1}\right)\right)}{2^{mp/2}\,|\boldsymbol{B}|^{(m+p+1)/2}\,\Gamma_p\left(\frac{m}{2}\right)}, \tag{3}$$

where $\Gamma_p\left(\frac{m}{2}\right) = \pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left(\frac{m}{2} - \frac{(i-1)}{2}\right)$ is the multivariate gamma function. For $p = 1$ the power inverse Wishart distribution is the distribution of Y^{-q} where $Y/\Psi \sim \chi^2_{(m)}$.

Theorem 1. *The function given in* (1) *is a density on the set of positive definite $p \times p$-matrices.*

As a preliminary for the proof, recall that any positive definite matrix $\boldsymbol{C}$ has a positive definite qth root given by $\boldsymbol{C}^{1/q} = \boldsymbol{V_C}\boldsymbol{\Delta}^{1/q}\boldsymbol{V_C}^{\top}$ where $\boldsymbol{V_C}$ is a orthonormal matrix diagonalizing $\boldsymbol{C}$, $\boldsymbol{\Delta} = \boldsymbol{V_C}^{\top}\boldsymbol{C}\boldsymbol{V_C}$ is the diagonal matrix of eigenvalues of $\boldsymbol{C}$ and $\boldsymbol{\Delta}^{1/q}$ is the diagonal matrix with the qth root of the eigenvalues of $\boldsymbol{C}$ in the diagonal (see, e.g. Mardia, Kent, and Bibby [9, Appendix A]).

Proof It follows from Deemer and Olkin [4, Theorem 3.7] that

$$\mathcal{W}^{-q}(\boldsymbol{B}|\boldsymbol{I}, m) = \mathcal{W}^{-q}\left(\boldsymbol{\Psi}^{1/2}\boldsymbol{B}\boldsymbol{\Psi}^{1/2}|\boldsymbol{\Psi}, m\right) \cdot \left|\boldsymbol{\Psi}^{1/2}\right|^{p+1}.$$

Thus it is sufficient to show that (1) is a density for $\boldsymbol{\Psi} = \boldsymbol{I}$.

Let $\boldsymbol{C}$ be an inverse Wishart-distributed matrix with parameters $\boldsymbol{I}$ and $m \geq p$, and consider the density of the distribution of the positive definite qth root $\boldsymbol{B} = \boldsymbol{C}^{1/q}$ of $\boldsymbol{C}$,

$$\mathcal{W}^{-1}(\boldsymbol{B}^q|\boldsymbol{I}, m) \cdot |J(\boldsymbol{B}^q, \boldsymbol{B})|\,,$$

where $J(\boldsymbol{B}^q, \boldsymbol{B})$ is the Jacobian matrix of the transformation $h(\boldsymbol{B}) = \boldsymbol{B}^q$ defined on the set of symmetric matrices. It follows from Magnus and Neudecker [8, p. 438 & Lemma 4.5(vi)] that

$$|J(\boldsymbol{B}^q, \boldsymbol{B})| = q^p\,|\boldsymbol{B}|^{q-1}\prod_{i<j}\frac{\lambda_i^q - \lambda_j^q}{\lambda_i - \lambda_j},$$

where $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$ are the eigenvalues of $\boldsymbol{B}$.The last term may be bounded from below as follows:

$$\begin{aligned}\frac{\lambda_i^q - \lambda_j^q}{\lambda_i - \lambda_j} &= \lambda_i^{q-1}\frac{1-(\lambda_j/\lambda_i)^q}{1-\lambda_j/\lambda_i} = \lambda_i^{q-1}\sum_{l=0}^{q-1}(\lambda_j/\lambda_i)^l \\ &\geq \frac{1}{\max_{l=0,\ldots,q-1}\binom{q-1}{l}}\sum_{l=0}^{q-1}\binom{q-1}{l}\lambda_j^l\lambda_i^{q-1-l} \\ &= \frac{(\lambda_i+\lambda_j)^{q-1}}{\max_{l=0,\ldots,q-1}\binom{q-1}{l}} \geq \frac{2^{(q-1)/2}}{\max_{l=0,\ldots,q-1}\binom{q-1}{l}}\cdot\sqrt{\lambda_i\lambda_j}^{\,q-1}.\end{aligned}$$

Thus

$$|J(\boldsymbol{B}^q, \boldsymbol{B})| \geq const \cdot |\boldsymbol{B}|^{q-1+(q-1)(p-1)/2}.$$

Hence the density of $\boldsymbol{B} = \boldsymbol{C}^{1/q}$ bounds

$$\begin{aligned}&\exp\left(-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{B}^{-q}\right)\right)|\boldsymbol{B}|^{-q/2(m+p+1)+q-1+(q-1)(p-1)/2} \\ &\qquad = \exp\left(-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{B}^{-q}\right)\right)|\boldsymbol{B}|^{-qm/2-p/2-1/2}\end{aligned}$$

up to a constant. It follows that (1) is integrable, and therefore it specifies a density. □

The next result, which describes the standard (i.e. $\boldsymbol{\Psi} = \boldsymbol{I}$) power inverse Wishart distribution, follows directly from Anderson [1, Theorem 13.3.4]:

Theorem 2. *Suppose B is a power inverse Wishart $(\boldsymbol{\Psi}, m, q)$-distributed $p \times p$-matrix and let $\lambda_1 > \lambda_2 > \ldots \lambda_p > 0$ denotes its eigenvalues and $\boldsymbol{V}$ the matrix containing its normalised eigenvectors chosen such that the first element of each column is non-negative.*

Then $(\lambda_1, \ldots, \lambda_p)$ and $\boldsymbol{V}$ are independent, the joint density of the eigenvalues is

$$g(\lambda_1, \ldots, \lambda_p) = \frac{\pi^{p^2/2}}{c_{m,q}\Gamma_p(p/2)}\cdot\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^p \lambda_i^{-q}\right)}{\prod_{i=1}^p \lambda_i^{(m+p+1)/2}}\cdot\prod_{i<j}(\lambda_i - \lambda_j),$$

and $\boldsymbol{V}$ has the conditional Haar invariant distribution (cf Anderson [1, Definition 13.3.1]).

The theorem says that the eigenvectors of a power inverse Wishart distributed matrix (including the inverse Wishart distribution) with $\boldsymbol{\Psi} = \boldsymbol{I}$ have the same distribution as the eigenvectors of a Wishart distributed matrix with the same matrix-parameter. Hence the distributions differ in how the eigenvalues are distributed.

It follows from Mardia et al. [9, Lemma 4.2.1] that the mode of the power inverse Wishart distribution is

$$\left(\frac{q}{qm+p+1}\right)^{1/q} \cdot \boldsymbol{\Psi}. \tag{4}$$

To compare the power inverse Wishart distribution to the inverse Wishart distribution, we look at the ratio

$$\frac{\mathcal{W}^{-q}(\boldsymbol{B}|\boldsymbol{I}, m_q)}{\mathcal{W}^{-1}(\boldsymbol{B}|\boldsymbol{I}, m_1)} = const \cdot \prod_{i=1}^{p} \exp\left(-\frac{1}{2}\left(\lambda_i^{-q} - \lambda_i^{-1}\right)\right) \lambda_i^{-(qm_q - m_1)/2}.$$

Here $\lambda_1, \ldots, \lambda_p$ denotes the eigenvalues of $\boldsymbol{B}$. We see that as any $\lambda_i \to 0$, this ratio goes to 0. Thus used as a prior for an unknown positive definite matrix, the general power inverse Wishart distribution gives smaller credibility to small eigenvalues, than does the usual inverse Wishart prior, and this effect gets stronger for larger values of q. The behaviour of the ratio as $\lambda_i \to \infty$ is determined by the parameters m_1 and m_q as well as by q: If $qm_q > m_1$, then the power inverse Wishart will penalise large eigenvalues harder, than the inverse Wishart does, whereas it will be more lenient if $qm_q < m_1$. If $qm_q = m_1$, then the ratio will approach a constant as $\lambda_i \to \infty$. Similar comments can be made in the case with a general $\boldsymbol{\Psi}$; in this case the eigenvalues $\lambda_1, \ldots, \lambda_p$ denotes the eigenvalues of $\boldsymbol{\Psi}^{-1/2}\boldsymbol{B}\boldsymbol{\Psi}^{-1/2}$. Thus $\boldsymbol{\Psi}$ is a "scaling parameter" and determines the position of the distribution, whereas q determines the tail behaviour at the "lower tail", and the product qm determines the upper tail behaviour.

We illustrate the tail behaviour in figure 1 for $p = 1$ and in figure 2 for $p = 2$ by plotting the ratios or the level curves of the ratios of the power inverse Wishart density to the inverse Wishart density for selected values of the parameters.

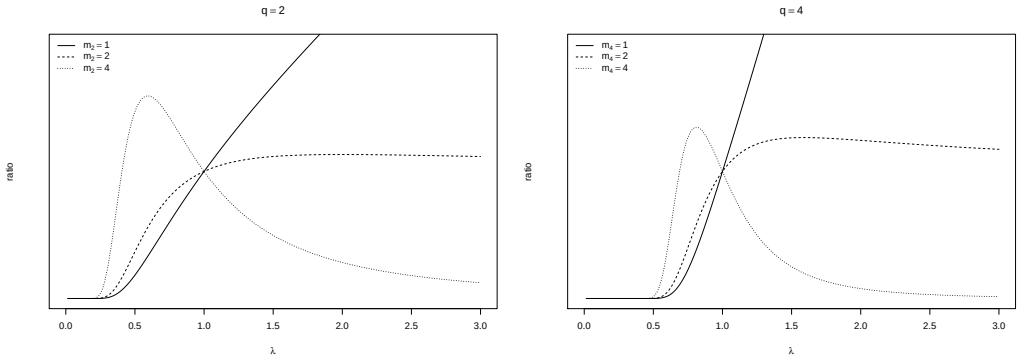

Figure 1: Ratio of the power inverse Wishart density to the inverse Wishart density for different values of the parameters. The left hand graph shows ratios for $q = 2$ and $m_1 = 4$, the right hand graph shows ratios for $q = 4$ and $m_1 = 8$. In both graphs ratios are given for $m_q = 1$, 2 and 4. The ratios have been normalized to take the same value at $\lambda = 1$.

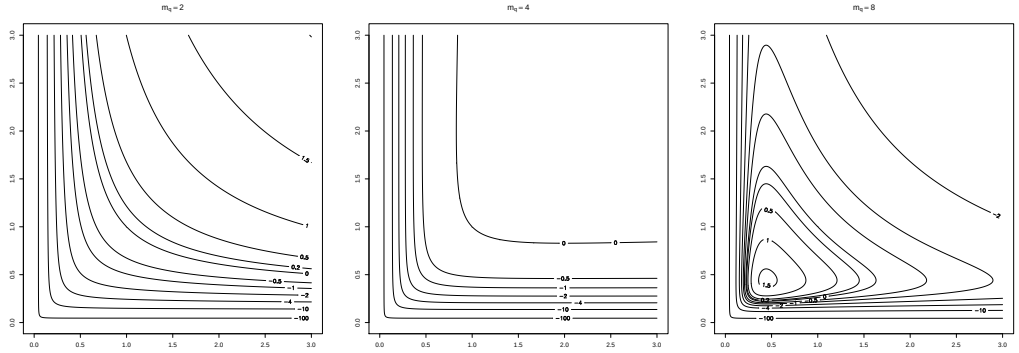

Figure 2: Ratio of the 2-power inverse Wishart density with $m_q = 2$, 4 and 8 to the inverse Wishart density with $m_1 = 8$. The ratios are normalized to take the value one at $(1, 1)$. The level curves are drawn at 10^c with the value of c $(-100, -10, -4, -2, -1, -0.5, 0, 0.2, 0.5, 1, 1.5, 2)$ denoted on the graphs.

3. Maximum A Posteriori Estimation

Consider a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$ of n independent and identically normally distributed p-dimensional random vectors, where both the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are unknown. The covariance matrix $\boldsymbol{\Sigma}$ is symmetric, and we will assume it to be positive definite. Put $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$ and let

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top$$

denote the empirical covariance matrix. Then the likelihood function is given by

$$\begin{aligned} L\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right) &= \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n (\boldsymbol{X}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu})\right)}{|\boldsymbol{\Sigma}|^{n/2}} \\ &= |\boldsymbol{\Sigma}|^{-n/2} \exp\left(-\frac{n}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right)\right) \\ &\qquad \cdot \exp\left(-\frac{n}{2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})\right). \end{aligned}$$

Provided that $n > p$, the likelihood function has a unique maximum at

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}, \qquad \hat{\boldsymbol{\Sigma}} = \boldsymbol{S}.$$

If $n \leq p$ the likelihood is unbounded, and in this case there is no maximum likelihood estimate (MLE). Of course $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ may still be used as estimators, but the properties of these estimators are typically poor. In many applications it may also be problematic that $\boldsymbol{S}$ is not positive definite. This is also the case when using methods such as principal components analysis or factor analysis. Even if the intention here is to reduce dimensionality, we would generally not want the reduction to be based simply on insufficient amounts of data leading to a singular covariance matrix. Moreover, if p is much larger than n, then the largest eigenvalue of $\boldsymbol{S}$ may severely overestimate the largest eigenvalue of $\boldsymbol{\Sigma}$ even if n is large (see section 5). One way of mending these problems is to put a prior distribution on the unknown parameters and use maximum a posteriori estimators. A standard choice of prior for $\boldsymbol{\Sigma}$ is the inverse Wishart distribution with parameters $(\boldsymbol{\Psi}, m)$. With an improper uniform prior on $\mathbb{R}^p$ for $\boldsymbol{\mu}$ this leads to MAP estimators given by

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}, \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{n+m+p+1}(n\boldsymbol{S} + \boldsymbol{\Psi}).$$

Without prior knowledge, a simple choice for the hyperparameter $\boldsymbol{\Psi}$ would be $\alpha \boldsymbol{I}$ for some α. This leads to an estimator of $\boldsymbol{\Sigma}$, which has the same eigenvectors as the MLE, but where the eigenvalues have been scaled down by $n/(n+m+p+1)$ and shifted upwards by $\alpha/(n+m+p+1)$. Thus, every eigenvalue of $\boldsymbol{S}$ is regularized in the same way regardless of its size. In some applications it may be more reasonable to apply different amounts of regularization depending on the size of the eigenvalue.

Instead of using an inverse Wishart prior for the unknown covariance matrix, $\boldsymbol{\Sigma}$, we propose to use a power inverse Wishart distribution as prior. Keeping the improper uniform prior for $\boldsymbol{\mu}$, the resulting posterior is given by

$$
\begin{aligned}
\pi\left(\boldsymbol{\Sigma}, \boldsymbol{\mu} | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right) &\propto L\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right) \cdot \mathcal{W}^{-q}(\boldsymbol{\Sigma} | \boldsymbol{\Psi}, m) \\
&\propto \frac{\exp\left(-\frac{n}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \boldsymbol{S}\right)\right) \cdot \exp\left(-\frac{n}{2}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})\right)}{|\boldsymbol{\Sigma}|^{1/2}} \\
&\qquad \cdot \frac{\exp\left(-\frac{1}{2} \operatorname{tr}\left(\left(\boldsymbol{\Psi}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{\Psi}^{-1/2}\right)^{-q}\right)\right)}{|\boldsymbol{\Sigma}|^{(n+p+qm)/2}} \\
&\propto \frac{\exp\left(-\frac{1}{2} \operatorname{tr}\left(n \boldsymbol{\Psi}^{-1/2} \boldsymbol{S} \boldsymbol{\Psi}^{-1/2} \cdot\left(\boldsymbol{\Psi}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{\Psi}^{-1/2}\right)^{-1}+\left(\boldsymbol{\Psi}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{\Psi}^{-1/2}\right)^{-q}\right)\right)}{\left|\boldsymbol{\Psi}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{\Psi}^{-1/2}\right|^{(n+p+qm+1)/2}} \\
&\qquad \cdot \exp\left(-\frac{n}{2}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})\right).
\end{aligned}
$$

Maximizing over $\boldsymbol{\mu}$ gives us $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$. In order to maximize over $\boldsymbol{\Sigma}$ we put $\boldsymbol{\mu} = \bar{\boldsymbol{X}}$, change parametrization to $\boldsymbol{\Upsilon} = \left(\boldsymbol{\Psi}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{\Psi}^{-1/2}\right)^{-1}$, and take logs to obtain

$$
\begin{aligned}
\boldsymbol{\Upsilon} \rightarrow l(\boldsymbol{\Upsilon}) &= \log \pi\left(\boldsymbol{\Sigma}, \boldsymbol{\mu} | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right) \\
&= -\frac{1}{2} \operatorname{tr}\left(n \boldsymbol{\Psi}^{-1/2} \boldsymbol{S} \boldsymbol{\Psi}^{-1/2} \cdot \boldsymbol{\Upsilon}+\boldsymbol{\Upsilon}^{q}\right)+\frac{n+p+qm+1}{2} \log |\boldsymbol{\Upsilon}| \\
&\qquad + \textit{const.}
\end{aligned}
\tag{5}
$$

Differentiating wrt. $\boldsymbol{\Upsilon}$ (see e.g. [8, Chapter 9]) gives us

$$\begin{aligned} dl(\boldsymbol{\Upsilon}) &= -\frac{1}{2}\operatorname{tr}\left(n\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}d\boldsymbol{\Upsilon} + q\boldsymbol{\Upsilon}^{q-1}d\boldsymbol{\Upsilon}\right) \\ &\qquad + \frac{n+p+qm+1}{2}\frac{1}{|\boldsymbol{\Upsilon}|}\operatorname{tr}\left(\boldsymbol{\Upsilon}^{-1}|\boldsymbol{\Upsilon}|\,d\boldsymbol{\Upsilon}\right) \\ &= -\frac{1}{2}\operatorname{tr}\left(\left(n\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2} + q\boldsymbol{\Upsilon}^{q-1} - (n+p+qm+1)\boldsymbol{\Upsilon}^{-1}\right)d\boldsymbol{\Upsilon}\right), \end{aligned}$$

which is 0, if

$$n\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2} + q\boldsymbol{\Upsilon}^{q-1} - (n+p+qm+1)\boldsymbol{\Upsilon}^{-1} = 0. \tag{6}$$

Differentiating again leads to

$$d^2l(\boldsymbol{\Upsilon}) = -\frac{1}{2}\operatorname{tr}\left(d\boldsymbol{\Upsilon}^{\top}\left(q(q-1)\boldsymbol{\Upsilon}^{q-2}d\boldsymbol{\Upsilon} + (n+p+qm+1)\boldsymbol{\Upsilon}^{-2}d\boldsymbol{\Upsilon}\right)\right),$$

so that the Hessian is negative definite. Moreover, by replacing $\boldsymbol{\Upsilon}$ in (5) by $t\boldsymbol{\Upsilon}$ it is easily shown that for any fixed $\boldsymbol{\Upsilon}$ the function $t \to l(t\boldsymbol{\Upsilon})$ tends to minus infinity as t tends to 0 or infinity. Thus we may conclude that $l(\boldsymbol{\Upsilon})$ has a unique maximizer, which solves (6) or equivalently

$$n\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}\cdot\boldsymbol{\Upsilon} + q\boldsymbol{\Upsilon}^{q} - (n+p+qm+1)\boldsymbol{I} = 0. \tag{7}$$

By transposing the terms of this equation, we see that any symmetric solution, $\hat{\boldsymbol{\Upsilon}}$, to this equation will commute with $\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}$. It follows that $\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}$ and $\hat{\boldsymbol{\Upsilon}}$ are diagonalized by the same orthonormal matrix (see [10, 1c(iii)]), and consequently the ith eigenvalue λ_i of $\hat{\boldsymbol{\Upsilon}}$ satisfies

$$q\lambda_i^q + n\lambda_i\left(\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}\right)\cdot\lambda_i - (n+p+qm+1) = 0, \tag{8}$$

where $\lambda_i(\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2})$ denotes the ith eigenvalue of $\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}$.

Theorem 3. *If we impose a power inverse Wishart prior distribution for* $\boldsymbol{\Sigma}$ *with parameters* $(m, \boldsymbol{\Psi}, q)$ *and an improper uniform prior of* $\boldsymbol{\mu}$*, then the maximum a posteriori estimator of* $\boldsymbol{\Sigma}$ *is*

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Psi}^{1/2}\boldsymbol{V}\hat{\boldsymbol{\Delta}}^{-1}\boldsymbol{V}^{\top}\boldsymbol{\Psi}^{1/2}, \tag{9}$$

where $\hat{\boldsymbol{\Delta}}$ *is a diagonal matrix with the unique positive solutions to the equations* (8) *in the diagonal, and* $\boldsymbol{V}$ *is an orthonormal matrix diagonalizing* $\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}$.

Proof The polynomial in (8),

$$\lambda \to q\lambda^q + n\lambda_i(\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}) \cdot \lambda - (n + p + qm + 1),$$

is negative for $\lambda = 0$ and goes to infinity as $\lambda \to \infty$. Furthermore, it is strictly increasing for $\lambda > 0$ so that (8) has exactly one positive solution. Hence $\hat{\boldsymbol{\Delta}}$ is well-defined. Moreover $\hat{\boldsymbol{\Upsilon}} = \boldsymbol{V}^\top \hat{\boldsymbol{\Delta}} \boldsymbol{V}$ clearly solves (7). It follows that

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Psi}^{1/2} \hat{\boldsymbol{\Upsilon}}^{-1} \boldsymbol{\Psi}^{1/2} = \boldsymbol{\Psi}^{1/2} \boldsymbol{V} \hat{\boldsymbol{\Delta}}^{-1} \boldsymbol{V}^\top \boldsymbol{\Psi}^{1/2}. \qquad \square$$

The positive solution of (8) is easily found numerically; we know that it is unique, and by Cauchy's bound [2] it is bounded by

$$1 + \max\left(n\lambda_i(\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}), n + p + qm + 1\right)/q.$$

Hence, we may solve (8) by a numerical method such as bisection. In the case $q = 2$, the eigenvalue equations (8) have closed form solutions

$$\lambda_i^{-1} = \frac{n}{2(n + p + 2m + 1)} \left(\lambda_i(\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}) + \sqrt{\lambda_i(\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2})^2 + 8\frac{n + p + 2m + 1}{n^2}} \right).$$

It follows that when $q = 2$, then

$$\hat{\boldsymbol{\Sigma}} = \frac{n}{2(n + p + 2m + 1)} \left(\boldsymbol{S} + \boldsymbol{\Psi}^{1/2} \left(\left(\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2} \right)^2 + 8\frac{n + p + 2m + 1}{n^2} \boldsymbol{I} \right)^{1/2} \boldsymbol{\Psi}^{1/2} \right),$$

which further simplifies to

$$\hat{\boldsymbol{\Sigma}} = \frac{n}{2(n + p + 2m + 1)} \left(\boldsymbol{S} + \left(\boldsymbol{S}^2 + 8\alpha^2 \frac{n + p + 2m + 1}{n^2} \boldsymbol{I} \right)^{1/2} \right), \quad (10)$$

when $\boldsymbol{\Psi} = \alpha \boldsymbol{I}$.

4. Regularization: Floor and shrinkage

In the previous section we derived the power inverse Wishart MAP, which includes the usual inverse Wishart MAP as a special case. In this section we will discuss and compare how the MAP estimators regularize the MLE. We will focus mainly on the case, where $\boldsymbol{\Psi}$ is a diagonal matrix, as this allows us to give some concrete expressions, but we will also comment on results for more general choices of $\boldsymbol{\Psi}$.

When $\boldsymbol{\Psi} = \alpha \boldsymbol{I}$ we may write

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{V} \hat{\boldsymbol{\Delta}} \boldsymbol{V}^\top, \tag{11}$$

where the orthonormal matrix $\boldsymbol{V}$ diagonalizes $\boldsymbol{S}$, and $\hat{\boldsymbol{\Delta}}$ is the diagonal matrix with diagonal elements given by the positive solutions to the equations

$$(n + p + qm + 1)\lambda_i^q - n\lambda_i\left(\boldsymbol{S}\right)\lambda_i^{q-1} - q\alpha^q = 0. \tag{12}$$

In this case, the MLE and the various MAP estimators all share the same eigenspaces, i.e. they are diagonalised by the same orthonormal matrix $\boldsymbol{V}$. The eigenvalues of the MAP estimators are the diagonal elements of $\hat{\boldsymbol{\Delta}}$ from (11), i.e. the solutions to the equations (12). Thus, the MAP estimators regularizes the eigenvalues of $\boldsymbol{S}$, but leave the eigenvectors unchanged. Hence their difference is, how the eigenvalues are regularized.

If λ is an eigenvalue of $\boldsymbol{S}$, then the corresponding eigenvalue for the inverse Wishart MAP estimator (3) is

$$\frac{1}{n + m + p + 1}(n\lambda + \alpha), \tag{13}$$

and for the 2-power inverse Wishart MAP (10) we get

$$\frac{n}{2(n + p + 2m + 1)}\left(\lambda + \sqrt{\lambda^2 + 8\frac{n + p + 2m + 1}{n^2}\alpha^2}\right). \tag{14}$$

Hence, both MAP estimators regularize the MLE by imposing a lower limit for the eigenvalues, which we denote the *floor*, and shrinking large eigenvalues by multiplying with a factor smaller than 1. In other words, both MAP estimators increase small eigenvalues and decrease large eigenvalues as compared to the MLE. We define the *shrinkage* as the limit of the regularized

eigenvalue divided by the corresponding unregularized eigenvalue as the latter tends to infinity. Thus, the shrinkage is the (asymptotic) scaling of large eigenvalues performed by the MAP estimator, whereas the floor is the lower limit for small eigenvalues imposed by the MAP estimator. The floor and the shrinkage factor both improve the estimation: The floor serves to make the estimator positive definite, whereas shrinking is beneficial for the estimation of the largest eigenvalues, as these tend to be overestimated, when p is not negligible compared to n (see also the following section).

For the inverse Wishart MAP, the floor and the shrinkage are

$$\frac{\alpha}{n+m+p+1} \quad \text{and} \quad \frac{n}{n+m+p+1}$$

respectively, whereas for the 2-power inverse Wishart the floor and shrinkage are

$$\frac{\sqrt{2}\,\alpha}{\sqrt{n+p+2m+1}} \quad \text{and} \quad \frac{n}{n+p+2m+1}$$

respectively. For general q the floor and shrinkage are

$$\alpha\left(\frac{q}{n+p+qm+1}\right)^{1/q} \quad \text{and} \quad \frac{n}{n+p+qm+1} \tag{15}$$

respectively. The floor follows directly from (12), which also shows that

$$\lambda_i \geq \frac{n}{n+p+qm+1}\lambda_i(\boldsymbol{S}). \tag{16}$$

Combining this with Cauchy's bound [2]

$$\lambda_i \leq 1 + \max\left(\alpha^q q, n\lambda_i(\boldsymbol{S})\right)/(n+p+qm+1), \tag{17}$$

we obtain the shrinkage given in (15) above.

The inverse Wishart MAP regularizes the eigenvalues by applying a linear function to the eigenvalues of $\boldsymbol{S}$; the power inverse Wishart MAP returns a strictly increasing and strictly convex function of the eigenvalues of $\boldsymbol{S}$. For $q = 2$ this follows directly from the expression (14). For general q, the Implicit Function Theorem gives us

$$\frac{d\lambda_i}{d\lambda_i(\boldsymbol{S})} = \frac{n}{q(n+p+qm+1) - n(q-1)\lambda_i(\boldsymbol{S})/\lambda_i}, \tag{18}$$

which is positive by (16), so that the function is increasing. Differentiating again we obtain

$$\frac{d^2\lambda_i}{d\lambda_i(\boldsymbol{S})^2} = \frac{n^2(q-1)}{(q(n+p+qm+1)-n(q-1)\lambda_i(\boldsymbol{S})/\lambda_i)^2}\frac{1}{\lambda_i}\left(1-\frac{\lambda_i(\boldsymbol{S})}{\lambda_i}\cdot\frac{d\lambda_i}{d\lambda_i(\boldsymbol{S})}\right),$$

which is positive, proving convexity. The convex regularization imposed by the power inverse Wishart prior has the effect that the difference between small eigenvalues after regularization is smaller than those between large eigenvalues. Thus the power inverse Wishart MAP regularizes eigenvalues differently depending on their sizes.

We also note that with the same floor and shrinkage, the eigenvalues of a power inverse Wishart MAP will always be smaller than the eigenvalues of the inverse Wishart MAP. Moreover, as the value of the derivative (18) at zero is a decreasing function of q, the eigenvalue of a power inverse Wishart MAP corresponding to any specific eigenvalue of $\boldsymbol{S}$ is decreasing as a function of the power q, when the floor and shrinkage are unchanged.

It is difficult to extend these results to the general case, where $\boldsymbol{\Psi}$ is not of the form $\alpha\boldsymbol{I}$, in a useful way. Clearly the results may be extended to results concerning the MAP estimator of $\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1/2}$ by replacing $\boldsymbol{S}$ with $\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}$, $\boldsymbol{\Sigma}$ with $\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1/2}$ and putting $\alpha = 1$. From this we see that the ith diagonal element of $\hat{\boldsymbol{\Delta}}^{-1}$ is larger than β and smaller than $\beta+\gamma\lambda_i(\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2})$, where β and γ are the floor and shrinkage respectively from (15) with $\alpha = 1$. Thus in the usual ordering of positive semi-definite matrices we have

$$\beta\boldsymbol{I} \leq \hat{\boldsymbol{\Delta}}^{-1} \leq \beta\boldsymbol{I} + \gamma\boldsymbol{\Delta},$$

where $\boldsymbol{\Delta}$ is the diagonal matrix with the eigenvalues of $\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}$ in the diagonal. From this we obtain

$$\beta\boldsymbol{\Psi} \leq \hat{\boldsymbol{\Sigma}} \leq \beta\boldsymbol{\Psi} + \gamma\boldsymbol{S}. \tag{19}$$

Thus in the general case, we may talk of a "matrix floor", $\beta\boldsymbol{\Psi}$, and also here there is a shrinkage effect, but the actual shrinkage may be smaller than the factor γ.

The inequalities in (19) has two trivial consequences that may be worth pointing out. The first is that similar inequalities hold for the diagonal elements of the matrices, i.e. for the estimated variances. The second consequence is that the MAP estimator has moments of all orders.

5. Asymptotic results

In a standard asymptotic set-up with $\boldsymbol{\Psi}$, m, and p fixed as n increases, the asymptotic behaviour of a power inverse Wishart MAP is the same as the asymptotic behaviour of the MLE.

Theorem 4. *Suppose that $\boldsymbol{\Psi}$, m, and p are fixed as n increases. Then the power inverse Wishart MAP $\hat{\boldsymbol{\Sigma}}$ and its eigenvalues have the same asymptotic distributions as the MLE $\boldsymbol{S}$.*

Proof First consider the case where $\boldsymbol{\Psi} = \alpha \boldsymbol{I}$. As the eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are bounded in probability by (17), it follows that $\hat{\boldsymbol{\Sigma}}^{q-1}$ is bounded in probability. Hence, re-writing (7) as

$$\left(\boldsymbol{S} - \frac{n+p+qm+1}{n} \hat{\boldsymbol{\Sigma}} \right) \hat{\boldsymbol{\Sigma}}^{q-1} = \frac{\alpha^q q}{n} \boldsymbol{I},$$

it follows that $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S} + o_P(1/\sqrt{n})$ and the result follows.

With a general fixed $\boldsymbol{\Psi}$, it follows that $\boldsymbol{\Psi}^{-1/2} \hat{\boldsymbol{\Sigma}} \boldsymbol{\Psi}^{-1/2} = \boldsymbol{\Psi}^{-1/2} \boldsymbol{S} \boldsymbol{\Psi}^{-1/2} + o_P(1/\sqrt{n})$, implying that also in this case the MAP estimator and the MLE have the same asymptotic distribution.

The results concerning the eigenvalues follow by continuous mapping. □

The densities of the limiting distributions in the case where $\boldsymbol{\Sigma} = \boldsymbol{I}$ are given in Anderson [1, Theorem 13.3.5].

As indicated in the introduction, our main interest is in estimating the covariance matrix in situations, where p is large compared to n. Assuming that the components of $\boldsymbol{X}_i$ are iid standard normal, and that both n and p increase such that $n/p \to \gamma \in [0; \infty]$, it is known that

$$\frac{\lambda_{\max}(\boldsymbol{S}) - \mu_{n,p}}{\sigma_{n,p}},$$

where $\lambda_{\max}(\boldsymbol{S})$ denotes the largest eigenvalue of $\boldsymbol{S}$, and $\mu_{n,p}$ and $\sigma_{n,p}$ are given by

$$\begin{aligned} \mu_{n,p} &= \left(1 + \sqrt{\frac{p}{n}} \right)^2, \\ \sigma_{n,p} &= \frac{\sqrt{n} + \sqrt{p}}{n} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{p}} \right)^{1/3}, \end{aligned} \tag{20}$$

converges in distribution to a Tracy-Widom distribution [6, 7]. For the MAP estimators we show the following result:

Theorem 5. *Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent, standard normally distributed random variables. Let $\lambda_{\max}^{(q)}$ denote the largest eigenvalue of the MAP estimator of $\boldsymbol{\Sigma}$ based on an power inverse Wishart prior with parameters $(\alpha \boldsymbol{I}, m, q)$. Then with $\mu_{n,p}$ and $\sigma_{n,p}$ as in* (20)*,*

$$\frac{\lambda_{\max}^{(q)} - \frac{n}{n+p+qm+1}\,\mu_{n,p}}{\frac{n}{n+p}\,\sigma_{n,p}}$$

converges in distribution to a Tracy-Widom distribution as $n, p \to \infty$ such that $n/p \to \gamma \in [0; \infty]$, and $m/p \to \kappa \in [1; \infty[$.

Proof The largest eigenvalue of the inverse Wishart MAP estimator is given by

$$\lambda_{\max}^{(1)} = \frac{1}{n+m+p+1}(n\lambda_{\max}(\boldsymbol{S}) + \alpha).$$

Consequently,

$$\frac{\lambda_{\max}^{(1)} - \frac{n}{n+m+p+1}\,\mu_{n,p}}{\frac{n}{n+p+m}\,\sigma_{n,p}} = \frac{\frac{n}{n+m+p+1}}{\frac{n}{n+p+m}}\frac{\lambda_{\max}(\boldsymbol{S}) - \mu_{n,p}}{\sigma_{n,p}} + \frac{\frac{\alpha}{n+m+p+1}}{\frac{n}{n+p+m}\sigma_{n,p}},$$

converges to a Tracy-Widom distribution, as $n\sigma_{n,p} \to \infty$.

A more indirect argument is needed for the general case. Recall that the eigenvalues solves (8), and that the derivative (18) is positive. This implies that $\lambda_{\max}^{(q)}$ solves (8) for $\lambda_i(\boldsymbol{S}) = \lambda_{\max}(\boldsymbol{S})$. Hence,

$$(n+p+qm+1)\left(\lambda_{\max}^{(q)} - \frac{n}{n+p+qm+1}\lambda_{\max}(\boldsymbol{S})\right)\left(\lambda_{\max}^{(q)}\right)^{q-1} = \alpha^q q. \quad (21)$$

Write

$$\frac{n}{n+p+qm+1}\lambda_{\max}(\boldsymbol{S}) = \frac{n}{n+p+qm+1}\,\sigma_{n,p}\cdot\frac{\lambda_{\max}(\boldsymbol{S}) - \mu_{n,p}}{\sigma_{n,p}} + \frac{n}{n+p+qm+1}\mu_{n,p},$$

and observe that the first term is $o_P(1)$ whereas the second term converges to a positive constant. Thus by the lower bound (16) it follows that $\lambda_{\max}^{(q)}$ is bounded away from 0 in probability. Consequently, we obtain

$$\frac{\lambda_{\max}^{(q)} - \frac{n}{n+p+qm+1}\lambda_{\max}(\boldsymbol{S})}{\sigma_{n,p}} = O_P\left(\frac{\sigma_{n,p}}{n+p+qm+1}\right) = o_P(1)$$

from (21), and hence

$$\frac{\lambda_{\max}^{(q)} - \frac{n}{n+p+qm+1}\mu_{n,p}}{\frac{n}{n+p+qm}\sigma_{n,p}} = \frac{\frac{n}{n+p+qm+1}}{\frac{n}{n+p+qm}} \cdot \frac{\lambda_{\max}(\boldsymbol{S}) - \mu_{n,p}}{\sigma_{n,p}} + o_P(1)$$

converges to a Tracy-Widom distribution. □

Remark. Recall that $m \geq p$ so that m must increase at least as fast as p. Hence in theorem 5, κ cannot be smaller than 1. A finite value of κ means that q increases at the same rate as p whereas $\kappa = \infty$ would mean that q increases at a faster rate. Note that our result does not include this scenario.

It follows from the proof of theorem 5 that

$$\lambda_{\max}^{(q)} \xrightarrow{P} \lim_{n,p\to\infty} \frac{n}{n+p+qm+1}\mu_{n,p} = 1 + \frac{2\sqrt{\gamma} - q\kappa}{1+\gamma+q\kappa},$$

where the last term is interpreted as 0, if γ equals ∞. Thus, the maximal asymptotic bias is smaller than 1. We note that the asymptotic bias of the largest eigenvalue of the power inverse Wishart MAP is bounded, whereas the asymptotic bias of the largest eigenvalue of the MLE is unbounded. In cases where $p < n$ (so that $\gamma > 1$) we may actually choose q and m such that the asymptotic bias is 0. Furthermore, the rate of convergence of the largest eigenvalue of the power inverse Wishart MAP is never slower than the rate of convergence of the largest eigenvalue of the MLE.

Remark. It is not obvious how to extend this result to the case, when $\boldsymbol{\Psi}$ is not of the form $\alpha\boldsymbol{I}$, since in this case the largest eigenvalue of the MAP estimator is not a simple function of the largest eigenvalue of the MLE. A related question is, what happens to the asymptotic results, when the covariance matrix of the underlying normally distributed data is $\boldsymbol{\Sigma}$ instead of

$\boldsymbol{I}$. In this case the largest eigenvalue of $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}$ has an asymptotic Tracy-Widom distribution. As

$$\lambda_{\max}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2})\lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{S}) \leq \lambda_{\max}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2})\lambda_{\max}(\boldsymbol{\Sigma}),$$

the asymptotic distribution of $\lambda_{\max}(\boldsymbol{S})$ depends on how the eigenvalues of $\boldsymbol{\Sigma}$ depends on p.

6. Simulations

To investigate the finite sample behaviour of our estimators we report on a small simulation study. We only consider the MLE, the usual inverse Wishart MAP and a power inverse Wishart MAP with $q = 2$. Both MAPs are based on priors with $\boldsymbol{\Psi} = \alpha\boldsymbol{I}$.

We consider two types of covariance matrices: The first is $\boldsymbol{\Sigma} = \boldsymbol{I}$, the second is a diagonal matrix with diagonal elements equal to

$$\boldsymbol{\Sigma}_{ii} = \begin{cases} \sigma^2\, i^{-0.7} & \text{for } 1 \leq i \leq p/10, \\ \sigma^2 \left(\frac{p}{10}\right)^{-0.6} i^{-0.1} & \text{for } p/10 < i \leq p, \end{cases} \tag{22}$$

which is illustrated in figure 3. Here there are a few large eigenvalues, but after a steep decrease the remaining eigenvalues are small and only decrease slowly. This covariance matrix is chosen to loosely mimic the behaviour of the eigenvalues in the real data example in the following section. We consider the behaviour of the MAP estimators under the quadratic loss function

$$L_2(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = \operatorname{tr}\left((\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})^\top\right).$$

The risk of the MLE and the inverse Wishart MAP can be calculated explicitly (see Appendix A), but the risk of the power inverse Wishart MAP cannot, so we rely on simulations. We will give results for three choices of p, namely 10, 50 and 100. For each value of p, we will use $n = p/2, p, 2p$ to investigate the behaviour in three different "asymptotic scenarios".

We first note that it is sufficient to consider diagonal matrices for $\boldsymbol{\Sigma}$: For any orthonormal matrix $\boldsymbol{V}$ we have

$$L_2(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = L_2(\boldsymbol{V}^\top\boldsymbol{\Sigma}\boldsymbol{V}, \boldsymbol{V}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{V}),$$

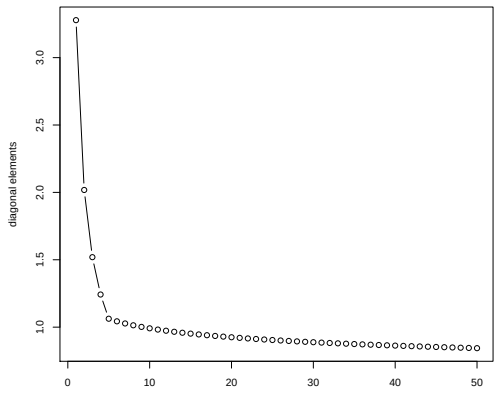

Figure 3: Eigenvalues of $\boldsymbol{\Sigma}$ for $p = 50$ used in the simulation.

Risk	$p = 10$	$p = 50$	$p = 100$
$n = p/2$	18	98	198
$n = p$	10	50	100
$n = 2p$	5.25	25.25	50.25

Table 1: Quadratic risk of the MLE.

and since

$$\boldsymbol{V}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{V} = \frac{n}{2(n+p+2m+1)} \Bigg(\boldsymbol{V}^\top \boldsymbol{S} \boldsymbol{V} + \left((\boldsymbol{V}^\top \boldsymbol{S} \boldsymbol{V})^2 + 8 \frac{n+p+2m+1}{n^2} \alpha \boldsymbol{I} \right)^{1/2} \Bigg)$$

for the 2-power inverse Wishart MAP (and with a similar result for the inverse Wishart MAP), the risks are left unchanged by rotations.

In our simulations, we choose σ^2 in (22), such that the risks of the MLE for given values of p and n are the same in the two examples; see table 1 for the values of these risks. We do not vary the variance parameter σ^2 in the simulations, because increasing σ^2 will give the same results as keeping it fixed while lowering the floor and scaling the resulting risks. Thus it is sufficient to vary the floor.

When comparing the two MAPs, the choice of hyperparameters is crucial: By choosing suitably different hyperparameters we can easily make one MAP looks superior to the other. To avoid this we try to choose the hyperparameters of the inverse Wishart prior so that the two MAP estimators have the same floor and the same overall amount of shrinkage. We believe that a reasonable comparison should use the same floor. However, if we use the same floor and the same shrinkage factor, then the regularization curve for the power inverse Wishart MAP (14) will be below the regularization curve for the inverse Wishart MAP (13) and our simulation results would be more a consequence of different amounts of shrinking rather than of the difference between the estimators. In order to circumvent this effect, we write (13) and (14) as $a'\lambda + b'$ and $a\lambda + a\sqrt{\lambda^2 + b}$ respectively. Here a and b are functions of the the chosen floor and shrinkage of the 2-power inverse Wishart MAP, and b' is just the common value of the floor. For chosen values of floor and shrinkage for the power inverse Wishart MAP, we choose a', such that

$$0 = \int_0^L \left(a\lambda + a\sqrt{\lambda^2 + b} - a'\lambda - b' \right) d\lambda$$

for a suitable value of L. Using $L = \infty$ leads to $a' = 2a$, i.e. the same shrinkage factor for the MAPs, so we need to choose a finite value of L. We choose L equal to the 99%-quantile in the distribution of the largest eigenvalue of the MLE. In this way the two MAPs has the same "average regularization" over the plausible range of observed eigenvalues.

The shrinkage factors of both MAPs are bounded by the fact that $m \geq p$. We use the maximal shrinkage factor for the power inverse Wishart MAP as well as factors 10% and 20% smaller. We also use three different values for the floor –0.8, 1, and 1.2– corresponding to the average value of the eigenvalues of $\boldsymbol{\Sigma}$ (to two decimal places for the matrix given by (22)) and values 20% smaller and larger.

The results based on 5,000 simulations are given in table 2 and 3. The differences between the two MAPs are small compared to the improvement over the MLE (see table 1). This is not unexpected. We have chosen the hyperparameters of the priors in order to make the MAP estimators as similar as possible, and all our simulations are in situations, where the MLE is not expected to work well. We see that choosing the floor equal to 1 typically leads to smaller risks. This is not surprising for the $\boldsymbol{\Sigma} = \boldsymbol{I}$ case, where all eigenvalues are equal to 1. Indeed, in this case it is optimal to use a floor equal to 1 ($\alpha = n + m + p + 1$) and shrink as much as possible ($m \to \infty$).

	floor	0.8			1			1.2		
(p,n)	shrink	1	0.9	0.8	1	0.9	0.8	1	0.9	0.8
(10,5)	q=1	*0.34*	*0.31*	*0.30*	0.26	0.20	0.15	0.97	0.87	0.79
	q=2	0.62	0.60	0.57	*0.10*	*0.08*	**0.07**	*0.21*	*0.22*	*0.23*
(10,10)	q=1	*0.40*	*0.33*	*0.27*	0.67	0.52	0.39	1.70	1.48	1.29
	q=2	0.44	0.43	0.41	*0.13*	*0.11*	**0.09**	*0.47*	*0.45*	*0.43*
(10,20)	q=1	0.72	0.53	0.37	1.51	1.17	0.90	3.01	2.56	2.16
	q=2	*0.30*	*0.27*	*0.25*	*0.32*	*0.26*	**0.20**	*1.00*	*0.91*	*0.84*
(50,25)	q=1	*1.40*	*1.37*	*1.36*	0.95	0.75	0.58	4.44	4.09	3.76
	q=2	3.12	2.99	2.86	*0.55*	*0.46*	**0.37**	*1.07*	*1.09*	*1.13*
(50,50)	q=1	*1.27*	*1.13*	*1.05*	2.08	1.64	1.26	6.77	6.05	5.39
	q=2	2.19	2.11	2.05	*0.68*	*0.56*	**0.45**	*2.32*	*2.23*	*2.15*
(50,100)	q=1	1.63	*1.22*	*0.92*	4.22	3.31	2.54	10.60	9.24	8.00
	q=2	*1.42*	1.31	1.23	*1.56*	*1.26*	**1.00**	*4.93*	*4.53*	*4.14*
(100,50)	q=1	*2.78*	*2.71*	*2.69*	1.95	1.54	1.19	9.00	8.27	7.60
	q=2	6.23	5.97	5.72	*1.10*	*0.92*	**0.75**	*2.15*	*2.19*	*2.26*
(100,100)	q=1	*2.52*	*2.24*	*2.07*	4.20	3.31	2.54	13.65	12.18	10.85
	q=2	4.38	4.22	4.09	*1.36*	*1.12*	**0.89**	*4.63*	*4.45*	*4.29*
(100,200)	q=1	3.26	*2.42*	*1.83*	8.45	6.64	5.09	21.26	18.54	16.06
	q=2	*2.83*	2.60	2.45	*3.11*	*2.53*	**2.00**	*9.85*	*9.05*	*8.28*

Table 2: Quadratic risk, $\boldsymbol{\Sigma} = \boldsymbol{I}$. Lines with $q = 1$ are for an inverse Wishart MAP, $q = 2$ for the power inverse Wishart MAP. The "shrink" is the factor multiplied onto the maximally possible shrinkage factor for the power inverse Wishart MAP. The smallest risk for each combination of (p, n) is given in bold; the smallest risk for each combination of (p, n) and floor and shrinkage is given in italics.

But it is also the case for the more realistic example, where most of the true eigenvalues are smaller than 1. Thus, it seems overall beneficial to overestimate small eigenvalues to some extent. On the other hand, as one would expect, it is also clear in our simulations that a floor that is "too small" is preferably to one that is "too large".

For the values used here, more shrinkage (smaller values of the shrinkage factor) generally leads to smaller risk, regardless of the floor for the values used here. Obviously, there will be a limit to this effect: If the floor is too low or too high, too much shrinking will lead to higher risks due to estimates that are too small or too large.

Overall the power inverse Wishart MAP performs better than the usual inverse Wishart MAP, when the floor is not too low. It should also be clear that we cannot conclude that the power inverse Wishart MAP is always

	floor	0.8			1			1.2		
(p,n)	shrink	1	0.9	0.8	1	0.9	0.8	1	0.9	0.8
(10,5)	q=1	*0.38*	*0.36*	*0.34*	0.31	0.25	0.20	1.02	0.92	0.84
	q=2	0.67	0.64	0.62	*0.15*	*0.13*	**0.12**	*0.26*	*0.27*	*0.27*
(10,10)	q=1	*0.44*	*0.36*	*0.31*	0.70	0.56	0.44	1.73	1.51	1.33
	q=2	0.48	0.47	0.46	*0.18*	*0.15*	**0.13**	*0.51*	*0.49*	*0.48*
(10,20)	q=1	0.74	0.55	0.41	1.52	1.20	0.93	3.04	2.59	2.20
	q=2	*0.33*	*0.30*	*0.29*	*0.35*	*0.29*	**0.24**	*1.04*	*0.95*	*0.88*
(50,25)	q=1	*7.33*	*7.42*	*7.53*	6.95	6.86	6.79	10.49	10.24	10.02
	q=2	9.24	9.19	9.15	*6.73*	*6.72*	**6.70**	*7.31*	*7.40*	*7.51*
(50,50)	q=1	*6.38*	*6.44*	*6.56*	7.28	7.02	6.84	12.02	11.50	11.02
	q=2	7.53	7.60	7.69	**6.12**	*6.14*	*6.18*	*7.83*	*7.89*	*7.95*
(50,100)	q=1	5.68	*5.55*	*5.56*	8.36	7.74	7.27	14.83	13.74	12.80
	q=2	*5.59*	5.72	5.91	*5.88*	**5.82**	**5.82**	*9.38*	*9.20*	*9.06*
(100,50)	q=1	*25.71*	*26.11*	*26.55*	25.06	25.09	25.18	32.24	31.94	31.71
	q=2	29.81	29.89	29.98	**24.89**	*25.01*	*25.16*	*26.08*	*26.42*	*26.81*
(100,100)	q=1	*22.25*	*22.76*	*23.39*	24.22	24.08	24.04	33.86	33.12	32.51
	q=2	24.72	25.19	25.72	**22.09**	*22.44*	*22.80*	*25.64*	*26.01*	*26.43*
(100,200)	q=1	18.81	19.15	*19.73*	24.39	23.71	23.31	37.48	35.84	34.49
	q=2	*18.32*	**19.14**	20.04	*19.21*	*19.61*	*20.14*	*26.43*	*26.57*	*26.82*

Table 3: Quadratic risk, $\mathbf{\Sigma}$ given by (22). Lines with $q = 1$ are for an inverse Wishart MAP, $q = 2$ for the power inverse Wishart MAP. The "shrink" is the factor multiplied onto the maximally possible shrinkage factor for the power inverse Wishart MAP. The smallest risk for each combination of (p, n) is given in bold; the smallest risk for each combination of (p, n) and floor and shrinkage is given in italics.

better than the usual inverse Wishart MAP. Along with the other hyperparameters, the power q must be chosen by the data analyst.

7. Application to real data

We consider the data set analysed by Shepstone, Rogers, Kirwan, and Silverman [11], who studied the intercondylar notch in human osteoarthritic and non-osteoarthritic femora. The authors considered 96 human femora from a large skeletal population. The femora were annotated by sex as well as distal eburnation. The available data is a sampling of a 2-dimensional spline curve approximation of the silhouette of the condyle in 50 arch length equidistant points normalised to start in (0,0) and end in (1,0).

We only consider a part of the data set, namely the 21 condyles with signs of distal eburnation. One of these (marked "2283R" in the data) differs markedly from the rest of the condyles (see figure 8), and we omit it from the estimation procedure. Later we will use the estimated covariance matrix to find a prediction of this condyle treating the middle part as missing. In this application, $n = 20$ whereas $p = 96$ (two times 50 minus the two end points, which are fixed).

In data like these, it would be natural to expect adjacent x (or y) coordinates to be highly correlated and distant x (y) coordinates to be less correlated, so we will let our choice of $\mathbf{\Psi}$ reflect this. The x and y coordinates may also be correlated, but we expect this correlation to be smaller, and we are not sure of its sign and put this part of the hyperparameter $\mathbf{\Psi}$ equal to 0. Also for simplicity, we assume variance homogeneity in our prior even though it is clear from the fact that the outlines of the notches have been "tied down" at the ends, that there will be less variation near the ends than in the middle. These considerations lead to $\mathbf{\Psi} = \alpha\mathbf{\Psi}_0$ with

$$\mathbf{\Psi}_0 = \begin{bmatrix} \mathbf{AR}(1)_\rho & \mathbf{0} \\ \mathbf{0} & \mathbf{AR}(1)_\rho \end{bmatrix},$$

where $\mathbf{AR}(1)_\rho$ is a correlation matrix for an AR(1)-process with parameter ρ, i.e. a matrix with (i, j)th element equal to $\rho^{|i-j|}$, and $\mathbf{0}$ is a matrix of 0s. Thus, we use the same correlation parameter for both x and y coordinates as well as assume variance homogeneity. This may be too simplistic, but without strong prior beliefs we prefer to keep $\mathbf{\Psi}$ simple. We use a prior with $q = 2$ and $m = p$; larger values of q and m leads to smaller shrinkage factors, and with p considerably larger than n we expect that this will give a sufficient

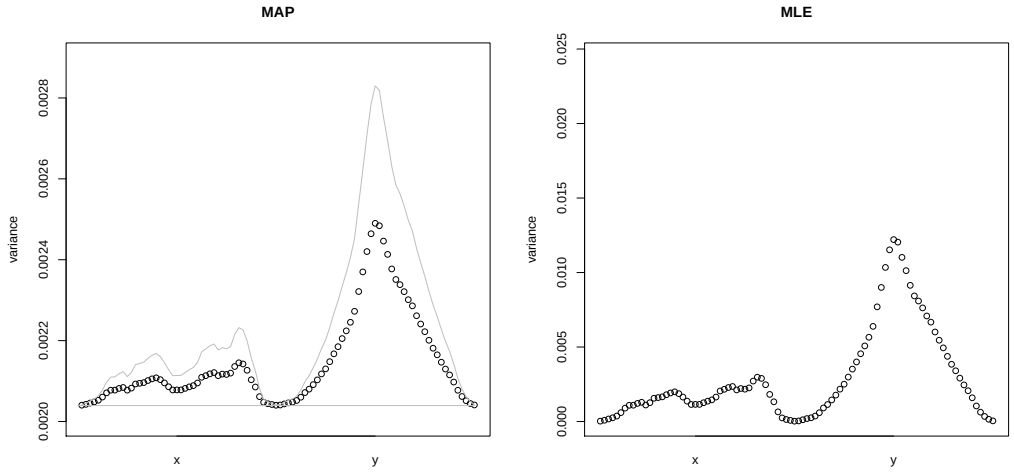

Figure 4: Estimated variances, the MLE on the right, the MAP on the left. The grey lines in the plot on the left are the bounds from (19).

amount of shrinkage. The values of $\rho\,(= 0.94)$ and $\alpha\,(= 0.02535)$ are chosen by predictive cross validation [5] using importance sampling.

Figure 4 shows how the estimated variances are lifted (by the floor) and shrunken, but also that the relative relationship between the variances are more or less unchanged. The MAP estimators of the large variances are much smaller than the MLEs, which of course is an effect of α and the shrinkage factor being fairly small; by (4) the prior mode is located at $0.0021\boldsymbol{\Psi}_0$. The smaller variances are lifted, and the averages of the estimated variances (the traces of the estimators) are not markedly different (0.0021 for the MAP and 0.0029 for the MLE).

Turning next to the estimated correlation matrix (figure 5), we see how the prior independence of x and y coordinates removes most of the correlation between x and y coordinates. The prior's AR(1)-structure is also evident in the correlations between x coordinates and between the y coordinates.

The eigenvalues and the first four eigenvectors of the MAP and the MLE are shown in figures 6 and 7. We see that the prior lifts the eigenvalues; only the largest eigenvalue is smaller when estimated by the MAP, than when it is estimated by the MLE. Note that the y-axis in figure 6 is logarithmic, so that the difference between the largest eigenvalues of the two estimators is rather big. The eigenvalues of the MAP estimator are pairwise similar. This

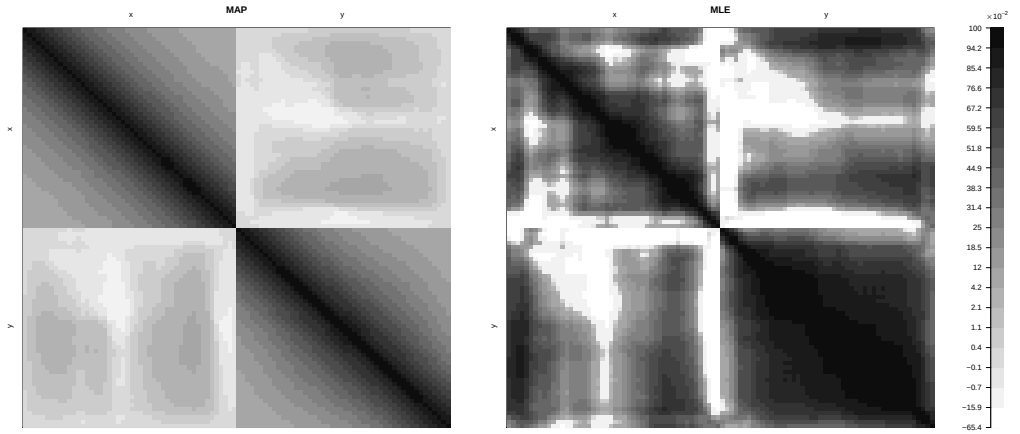

Figure 5: Estimated correlation matrices; MLE on the right, MAP on the left. The grey tone-bands are based on the 5%-, 10%-, ..., 95%-quantiles of the elements of the two estimators.

is probably an effect of the block-diagonal $\mathbf{\Psi}$; it tends to split the variation into a part mostly related to the x-coordinates and a part mostly related to the y-coordinates. This is also what we see from figure 7. Indeed it seems that the sinusoidal-looking eigenvectors of AR(1)-correlation matrices and the block diagonal form have a dominant effect on the resulting MAP estimator.

Any application of MAP estimation is a compromise between the data and the prior: We wish to balance the information provided by the data with the stability introduced by the prior. It is not surprising that the prior has a large effect in this example. Even if we suspect that the true covariance matrix is more complicated, there hardly is any information in the data to help us discover it. The size of dataset is very small compared to the dimension of the unknown covariance matrix, so the shrinkage factor is quite small, and $\mathbf{\Psi}$ has a lot of weight in the resulting estimator. Though this is the intended effect of MAP estimation, it also means that the prior should be chosen carefully. In this example we have used a very simple choice of $\mathbf{\Psi}$. More complicated choices may be considered: Different variances for x- and y-coordinates, as well as correlation between x and y-coordinates are easily implemented in the estimator. However, choosing the values of the hyperparameters is more complicated. Our solution to this problem is basically a grid search, and

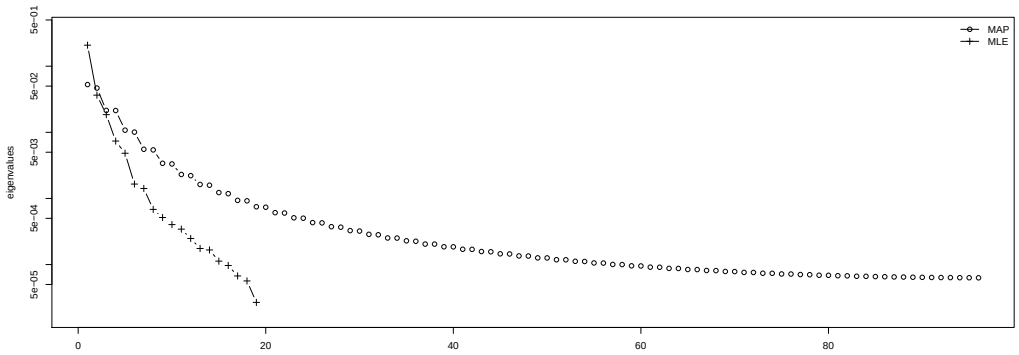

Figure 6: Eigenvalues of the MLE ('+') and the MAP ('∘'); note that the y-axis is logarithmic.

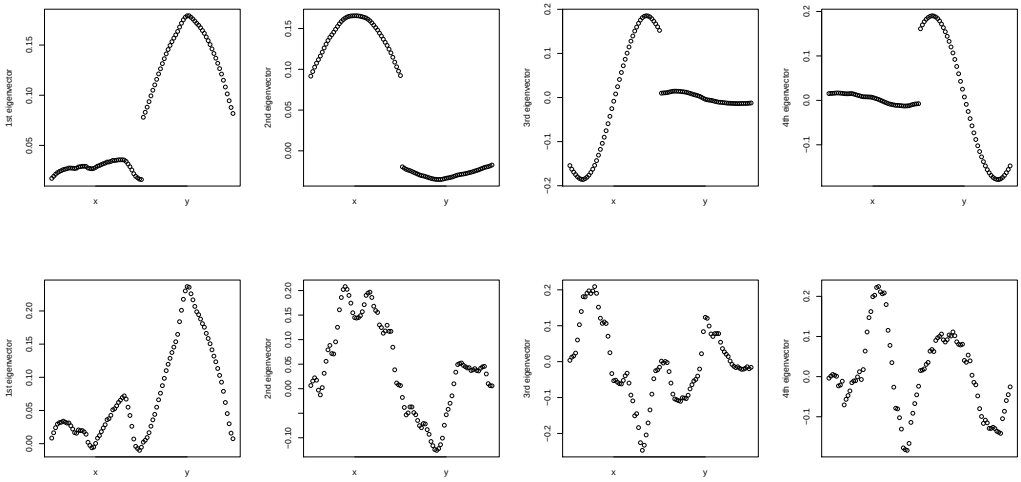

Figure 7: First 4 eigenfunctions of the estimators. The top row is the MAP, the bottom row is the MLE.

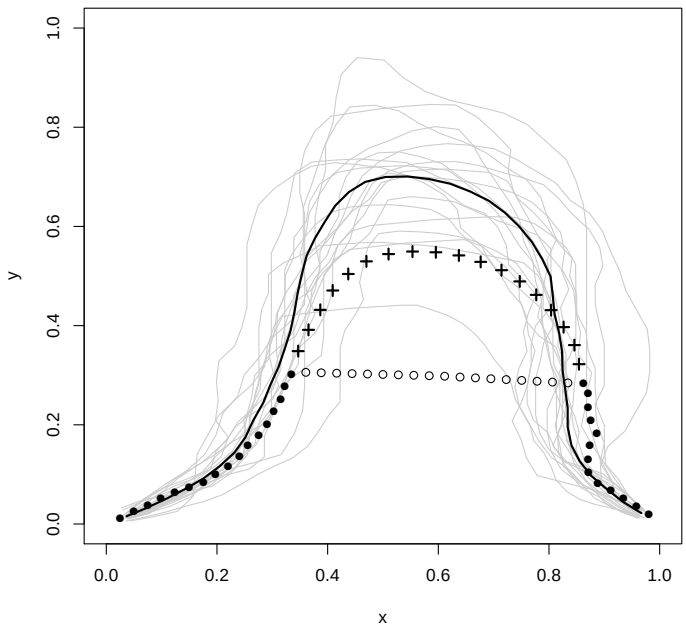

Figure 8: Prediction of the middle part of notch "2283R". The full line is the mean shape of the notches, the grey lines the observed notches. The circles and bullets outline notch "2283R" as it is in the data. We interpret the bullets as observed and the circles as missing. The prediction of the missing part is given by the pluses.

the more parameters that need to be chosen, the longer the computation time. For this reason, we will not attempt a more complicated prior for this example.

As mentioned at the beginning of this section the condyle "2283R" differs radically from the rest. As seen in figure 8, where it is represented by circles and bullets, it seems to have had its middle part "cut off", when compared to the other condyles in the dataset (grey curves in the figure). As an illustration we pretend that the middle part (the part of the condyle represented by the circles) are missing data and try to predict it. The usual EBLUP formula (see e.g. Anderson [1, p. 37]) based on the MLE breaks down; there are 30 observed points (the bullets in the figure), so with only 20 fully observed condyles the covariance matrix corresponding to the observed part of "2283R" is singular and cannot be inverted. The MAP, on the other hand, is

regular, and when using this in the formula, we obtain the prediction given in figure 8 by the pluses.

8. Conclusion

In this paper we have introduced a new class of distributions –the power inverse Wishart distributions– on the set of positive definite matrices. Used as priors for unknown covariance matrices of multivariate Gaussian data, they lead to easily calculable maximum a posteriori estimators. Our simulations suggest that the MAP estimators perform better than the MLE in terms of overall quadratic risk. We have derived some asymptotic properties of these estimators and have seen that these are as good as or in some situations even better than those of the MLE.

As we have seen in sections 6 and 7 the choice of prior influences the MAP estimator. Obviously, if this was not the case, there would be little reason for using the MAP estimator. On the other hand, it also means that the prior should be chosen carefully. In section 7 we chose the form of the prior mode based on prior beliefs but the values of it was determined by cross validation. Our implementation of this cross validation is too computationally demanding to allow a further investigation of its properties, so it is difficult to know if this is in any sense optimal. Clearly, this is an area that requires additional work.

It is quite easy to extend our results (except theorem 1) to improper priors with $m < p$; in theorem 5 this would allow κ to be any non-negative real. By allowing improper priors, we could obtain a MAP estimator in our example with less shrinkage than the one we have used. It is less obvious whether our results can be extended to values of q that are not positive integers, as many of our arguments rely on q being a positive integer.

We hope that the additional flexibility provided by the power inverse Wishart MAP will prove to be useful when estimating large covariance matrices based on limited amounts of data.

Acknowledgments

We gratefully acknowledge the funding from the Danish Research Foundation (Den Danske Forskningsfond) supporting this work, and Bo Markussen, University of Copenhagen, for useful comments.

Appendix A. Quadratic risk

For estimators of the form

$$\hat{\boldsymbol{\Sigma}} = a\boldsymbol{S} + b\boldsymbol{I}$$

such as the MLE and the inverse Wishart MAP estimator, the expected quadratic risk is

$$\begin{aligned} E[L_2(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})] &= \operatorname{tr}\left(E\left[(a\boldsymbol{S} + b\boldsymbol{I} - \boldsymbol{\Sigma})^2\right]\right) \\ &= a^2\left(\frac{n-1}{n}\sum_{i=1}^{p}\sum_{j=1}^{p}\Sigma_{i,j}^2 + \frac{n-1}{n^2}\left(\sum_{i=1}^{p}\Sigma_{i,i}\right)^2\right) \\ &\quad + 2a\frac{n-1}{n}\left(b\operatorname{tr}\left(\boldsymbol{\Sigma} - \operatorname{tr}\left(\boldsymbol{\Sigma}^2\right)\right)\right) \\ &\quad + b^2 p - 2b\operatorname{tr}\left(\boldsymbol{\Sigma}\right) + \operatorname{tr}\left(\boldsymbol{\Sigma}^2\right) \end{aligned}$$

as S is Wishart distributed with parameters $(n-1, \boldsymbol{\Sigma}/n)$.

The expression for $E[L_2(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})]$ is a convex polynomial of (a, b) of degree two and thus has a minimal value. Thus, there are unique optimal values of the floor and the shrinkage for the inverse Wishart MAP (for a given $\boldsymbol{\Sigma}$), but there are also choices that will lead to inverse Wishart MAPs with larger risks than the MLE.

References

[1] Anderson, T. W., 2003. An Introduction to Multivariate Statistical Analysis, 3rd Edition. Wiley, New York.

[2] Cauchy, A. L., 1829. Exercises de mathématique. In: Oeuvres (2). Vol. 9. , p. 122.

[3] Crimi, A., Lillholm, M., Mads Nielsen, A. G., de Bruijne, M., Dam, E. B., Sporring, J., 2010. Bayes estimation of linear shape variation with application to vertebra and cartilage modeling, (submitted).

[4] Deemer, W. L., Olkin, I., 1951. The jacobians of certain matrix transformations useful in multivariate analysis: Based on lectures of P. L. Hsu at the University of North Carolina. Biometrika 38, 345–367.

[5] Gelfand, A. E., Dey, D. K., Chang, H., 1992. Model determination using predictive distributions with implementation via sampling-based metods. In: Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M. (Eds.), Bayesian Statistics. Vol. 4. Oxford University Press, Oxford, UK, pp. 147–167.

[6] Johnstone, I. M., 2001. On the distribution of the largest eigenvalue in principal componentes anlysis. Ann Statist 29 (2), 295–327.

[7] Karoui, N. E., 2008. On the largest eigenvalue of Wishart matrices with identity covariance when n, p and $p/n \to \infty$. Tech. rep., arXiv:math/0309355v1.

[8] Magnus, J. R., Neudecker, H., 1980. The elimination matrix: Some lemmas and applications. SIAM J. Alg. Disc. Meth. 1 (4), 422–449.

[9] Mardia, K. V., Kent, J. T., Bibby, J. M., 1979. Multivariate Analysis. Academic press, London, UK.

[10] Rao, C. R., 1973. Linear Statistical Inference and Its Applications. Wiley, New York.

[11] Shepstone, L., Rogers, J., Kirwan, J., Silverman, B., 2001. Shape of the intercondylar notch of the human femur: a comparison of osteoarthritic and non-osteoarthritic bones from a skeletal sample. Ann Rheumati Dis 60, 968–973.